"FEELnc: A tool for Long non-coding RNA annotation and its application to the dog transcriptome" by Wucher V. *et al.*

This document aims at providing supplemental information on 1) FEELnc program predictors, 2) the material, input files, exact command-lines and R scripts to make figures used in the program benchmarking, 3) a guideline for FEELnc, 4) the mapping and transcript models reconstruction, 5) rules to build the new canine annotation isoforms (CanFam3.1-plus) and 6) transcript and gene biotypes definitions.

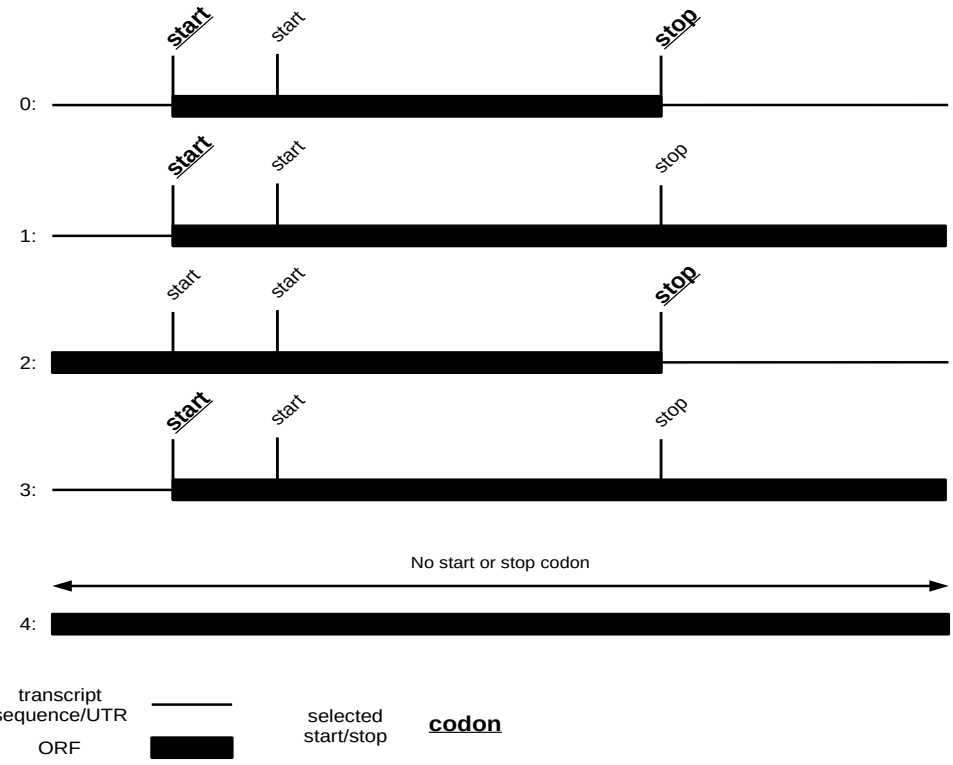
In addition, supplementary tables and figures are also inserted at the end of the document.

1- FEELnc ORF Type definition

FEELnc computes five types of ORFs depending on the presence/absence of one or both start and stop codons and the longest ORF is selected. The 5 ORF types are illustrated by the schema below and are defined as:

- '0': ORFs with start and stop codon;
- '1': same as '0' + ORFs with only a start codon;
- '2': same as '0' + ORFs with only a stop codon;
- '3': same as '0' + ORFs with a start or a stop (see '1' and '2');
- '4': same as '3' but if no ORF is found, take the whole input sequence as ORF.

Note that if the CDS is annotated in the input reference .GTF file, then the CDS is automatically selected as the longest ORF.



Schema of the 5 ORF types annotated in FEELnc. The start and/or stop codons selected by FEELnc is/are underlined.

2- Benchmarking programs and data availability

For each tool (except for PhyloCSF, see below) tested on the GENCODE datasets (human and mouse), a protocol is available containing all command-lines and links to the data used in order to replicate the analysis (note that firefox or chrome browsers are preferred for readability):

- Human:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/human/README_human

- Mouse:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/mouse/README_mouse

- Without lncRNAs learning dataset:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/without_lncRNAs_learning/README_without_lncRNAs_learning

- R scripts to generate paper figures:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/figure_feelnc.tgz

In order to run PhyloCSF, we first downloaded multiple genome alignment provided by the UCSC browser [1]. We chose the human (GRCh38) alignment against 19 mammalian genomes (20way). Then, using the Stitch Gene blocks tool in Galaxy [2,3], we extracted from the multispecies alignment the alignment for each sequences in our *human testing dataset* (HT, see main article). Then, we used the PhyloCSF Docker container [4,5] with the PhyloCSF 58 mammals model and the sequence alignments of the 15 species in the 20way alignment who are also in the 58 mammals PhyloCSF model. For each sequence, the gaps in human sequence were removed. The ORFs have been predicted for the three frames and an ORF is defined as either: i) the sequence between two stop codons or ii) between the beginning of the sequence and a stop codon or iii) between a stop codon and the end of the sequence. For a sequence, the best score among all predicted ORFs gives his coding probability.

3- FEELnc guidelines

Here, we provide two guidelines describing how to use FEELnc for typical analyses **with** and **without** a reference annotation. Note that:

- The step '1' is **not part of** FEELnc but could usually be done by transcriptome reconstruction programs either genome-guided (Cufflinks, StringTie…) or *de novo* assembly tools (Trinity, KisSplice…).
- Each of these steps can be run independently.
- The **FEELnc$_{codpot}$** step can be run using either **GTF** or **FASTA** files, that is why it can be can used **with** or **without** a reference genome. Even with a reference genome, it can be run using **FASTA** format. In contrary, the two other modules, **FEELnc$_{filter}$** and **FEELnc$_{classifier}$** do require **GTF** files as input and so, implicitly, a reference genome.

Minimal command line examples are provided for each step (except '1').

In the case of an annotation **with** a reference genome:

1. **Transcript models reconstruction**, the first step is to reconstruct transcript model from RNA-seq experiment with **dedicated** tool as Cufflinks or Stringtie. This will lead to a **GTF** file containing the coordinates of new potential transcripts, later denoted as the **transcript models file** (in **GTF**).

2. **FEELnc$_{filter}$** flags potential transcripts of the transcript models file which overlap a reference annotation file (in GTF format). This leads to a file with transcript models that do not overlap the reference annotation, denoted later as **candidate models file** (in **GTF**).

```
FEELnc_filter.pl -i transcript_model.gtf -a reference_annotation.gtf
> candidate_model.gtf
```

3. **FEELnc$_{codpot}$** calculates a coding potential score and infer the biotype, coding/non-coding, of the models in the candidate lncRNA transcript models file. It can be used with the **candidate models file (**in **GTF**) and the genome sequence to extract the FASTA sequences from it. Concerning the learning step of FEELnc$_{codpot}$, three strategies can be applied with respect to the availability of a learning lncRNA file:

   (a) Annotated lncRNAs are available: FEELnc$_{codpot}$ will use them to learn is random forest model.

```
FEELnc_codpot.pl -i candidate_model.gtf -a reference_mrna.gtf -
l reference_lncrna.gtf -g genome.fasta
```

   (b) No annotated lncRNAs are available but annotated lncRNAs from a closest species (< ~100 Myr) are available: FEELnc$_{codpot}$ will use them to learn its random forest model.

```
FEELnc_codpot.pl -i candidate_model.gtf -a reference_mrna.gtf -
l close_species_lncrna.fasta -g genome.fasta
```

   (c) No annotated lncRNAs are available and no annotated lncRNAs from a closest species are available: FEELnc$_{codpot}$ will shuffle the set of mRNAs in order to simulate a set of lncRNAs.

```
FEELnc_codpot.pl -i candidate_model.gtf -a reference_mrna.gtf -
g genome.fasta --mode=shuffle
```

   Once a strategy have been chosen, the user can use two distinct threshold methods:

   i. Use the **automatic** threshold (default).

```
No option needs to be specified.
```

   ii. Provide **two specificity** thresholds. Example with a 0.95 specificity threshold for mRNAs and lncRNAs for strategy (a):

```
FEELnc_codpot.pl -i candidate_model.gtf -a reference_mrna.gtf -
l reference_lncrna.gtf -g genome.fasta --spethres=0.95,0.95
```

This leads to at least **two transcripts files** (in **GTF**), a **lncRNAs file** and a **mRNAs file**. In case of the second strategy, a **TUCps file** (in **GTF**) is provided.

4. **FEELnc$_{classifier}$** classifies the lncRNAs file regarding closest transcripts or transcripts which overlap in a window. This classification is done with respect to transcripts included in a user provide reference annotation file. This can be done with all biotypes. This lead to a **column-separated file** with each line representing a lncRNA, a transcript, the distance/overlap between them and the class of this lncRNA regarding this transcript.

```
FEELnc_classifier.pl -i new_lncRNA.gtf -a reference_annotation.gtf >
new_lncRNA_classes.txt
```

In case of an annotation **without** a reference genome:

1. **Transcript models reconstruction**, the first step is to reconstruct transcript model from RNA-seq experiment with **dedicated** tool as Trinity or KisSplice. This will lead to a **FASTA** file containing sequences of new transcripts, later denoted as the **models file**.
2. **FEELnc$_{codpot}$** calculates a coding potential score and infer the biotype, coding/non-coding, of the assembled transcript models. It can be used with the **models file (**in **FASTA**). Concerning the learning step of FEELnc$_{codpot}$, three strategies can be applied with respect to the availability of a learning lncRNA file:
   (a) Annotated lncRNAs are available: FEELnc$_{codpot}$ will use them to learn is random forest model.

```
FEELnc_codpot.pl -i model.fasta -a reference_mrna.fasta -l
reference_lncrna.fasta
```

   (b) No annotated lncRNAs are available but annotated lncRNAs from a closest species (< ~100 Myr) are available: FEELnc$_{codpot}$ will use them to learn is random forest model.

```
FEELnc_codpot.pl -i model.fasta -a reference_mrna.fasta -l
close_species_lncrna.fasta
```

(c) No annotated lncRNAs are available and no annotated lncRNAs from a closest species are available: FEELnc$_{codpot}$ will shuffle the set of mRNAs in order to simulate a set of lncRNAs.

```
FEELnc_codpot.pl -i model.fasta -a reference_mrna.fasta --mode=shuffle
```

Once a strategy have been chosen, the user can used two distinct threshold methods:

i. Use the **automatic** threshold (default).

```
No option needs to be specified.
```

ii. Provide **two specificity** thresholds. Example with a 0.95 specificity threshold for mRNAs and lncRNAs for strategy (a):

```
FEELnc_codpot.pl -i model.fasta -a reference_mrna.fasta -l reference_lncrna.fasta --spethres=0.95,0.95
```

This leads to at least **two transcripts files** (in **FASTA**), a **lncRNAs file** and a **mRNAs file**. In case of the second strategy, a **TUCps file** (in **FASTA**) is provided.

4- Reads mapping and transcript models reconstruction of canine RNA-seq samples

The mapping of the reads has been made using STAR v2.5.0a and the models reconstruction with Cufflinks v2.2.1. Details on command lines are available at the following URL:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/dog_reannotation/README_mapping_reconstruction

Note: in order to get the full compatibility between STAR and Cufflinks, the option '`--alignEndsType EndToEnd`' is needed in STAR.

5- Canine extended coding potential isoforms

For the prediction of the transcript biotypes in the extended dog annotation (CanFam3.1-plus), command lines, parameters and input data are freely available at the following URL:

http://tools.genouest.org/data/tderrien/FEELnc_article_supplementary/dog_reannotation/README_dog_reannotation

From the set of transcripts that overlapped the reference annotation (CanFam3.1), a transcript is considered as a new isoform if: neither all exons are included at 100% in the previous annotation, nor all introns are included at 100% in the previous annotation. After the identification of new isoforms, some of the remaining could overlap two already annotated

genes in CanFam3.1 annotation resulting in the merging of these two genes. In order to parsimoniously merge genes from the reference annotation, we only kept new merging isoforms when their biotypes correspond to the biotypes of each transcript of the reference genes.

6- CanFam3.1-plus transcript and gene biotypes

Using the FEELnc$_{classifier}$ module and a set of homemade rules, the lncRNAs from CanFam3.1-plus have been classified according to the 5 following classes:

- Genic long non-coding RNA (glncRNA): a lncRNA who is sense exonic or intronic of an mRNA (note that this only concerns new isoforms of already annotated transcripts);
- Host long non-coding RNA (hlncRNA): a lncRNA who is sense exonic or intronic of a small non-coding RNA (sncRNA) (snoRNA, miRNA, rRNA, snRNA, etc);
- Messenger RNA antisense (mRNA antisense): a lncRNA who is antisense exonic or intronic of an mRNA;
- Non-coding RNA antisense (ncRNA antisense): a lncRNA who is antisense exonic or intronic of a sncRNA;
- Long intergenic non-coding RNA (lincRNA): a lncRNA who is neither sense or antisense of an mRNA or a sncRNA.

In order to define the gene biotypes for all transcript biotypes belonging to this gene, we check every transcript biotypes and select the gene biotype with respect to the following hierarchy:

1. Protein-coding;
2. Antisense mRNA;
3. Antisense ncRNA;
4. glncRNA;
5. hlncRNA;
6. lincRNA;
7. Processed pseudogene, pseudogene and misc_RNA;

8. miRNA, snoRNA, snRNA and rRNA;
9. TUCp.

Supplemental Tables and Figures

Species		mRNA transcripts	lncRNA transcripts
Nematode	Caenorhabditis elegans	30,939	3,271
Chicken	Gallus gallus	16,354	13,085
Chimpanzee	Pan troglodytes	19,907	18,604
Cow	Bos taurus	22,118	23,696
Fly	Drosophila melanogaster	30,362	54,819
Gorilla	Gorilla gorilla	27,473	20,785
Opossum	Monodelphis domestica	22,310	21,014
Orangutan	Pongo abelii	21,414	15,601
Platypus	Ornithorhynchus anatinus	23,584	11,518
Rat	Rattus norvegicus	28,635	29,070
Rhesus	Macaca mulatta	36,384	9,325
Zebrafish	Danio rerio	44,052	5,014
Arabidopsis	Arabidopsis thaliana	12,956	3,853

Supplementary Table 1: Number of available mRNA and lncRNA transcripts. All mRNA transcripts come from EnsEMBL, except for *Arabidopsis thaliana* where the mRNAs come from the TAIR database (https://www.arabidopsis.org/). All lncRNA transcripts come from the NONCODE database version 2016 (http://noncode.org/).

Organ	Dog breed	Total reads	Mapped reads	Transcripts by Cufflinks	Sample_owner
Adrenal_gland	Bernese Mountain Dog	54,776,586	88.97%	135,927	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Cerebellum	Belgian Shepherd	52,776,728	91.70%	147,602	Dr. M. Fredholm: University of Copenhagen, Denmark
Cerebellum	Great Swiss Mountain Dog	44,902,865	90.94%	145,034	Dr. M. Fredholm: University of Copenhagen, Denmark

Cortex	Belgian Shepherd	41,319,413	91.50%	137,780	Dr. M. Fredholm: University of Copenhagen, Denmark
Gut_colon	Bernese Mountain Dog	52,310,396	89.55%	135,462	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Hair_follicule	Labrador	45,694,722	82.67%	137,922	Dr. T. Leeb: University Bern, Switzerland
Jejunum	Labrador	50,569,866	86.91%	129,404	Dr. H Fieten: Utrecht University, Netherlands
Keratinocyte	Beagle	54,482,221	88.37%	129,950	Dr. T. Leeb: University Bern, Switzerland
Mammary_gland	Great Swiss Mountain Dog	44,349,725	87.43%	137,290	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Nose*	Labrador	58,842,156	87.66%	143,488	Dr. T. Leeb: University Bern, Switzerland
Nose*	Labrador	68,181,155	88.90%	168,477	Dr. T. Leeb: University Bern, Switzerland
Nose*	Labrador	69,193,538	84.34%	143,209	Dr. T. Leeb: University Bern, Switzerland
Olfactory_bulb	Great Swiss Mountain Dog	45,799,491	88.44%	132,777	Dr C. André: University Rennes1 –

					IGDR-CNRS Rennes, France
Pancreas	Belgian Shepherd	47,171,936	79.01%	123,558	Dr. M. Fredholm: University of Copenhagen, Denmark
Retina	Border Collie	50,480,134	83.40%	136,797	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Skin	Beagle	52,275,710	86.43%	131,191	Dr. T. Leeb: University Bern, Switzerland
Skin	Great Swiss Mountain Dog	49,080,979	84.98%	133,627	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Spinal_cord	Great Swiss Mountain Dog	46,844,306	89.92%	132,420	Dr C. André: University Rennes1 – IGDR-CNRS Rennes, France
Spleen	Belgian Shepherd	49,604,583	82.57%	139,643	Dr. M. Fredholm: University of Copenhagen, Denmark
Thymus	Saluki	51,079,197	87.96%	135,265	Dr. H. Lohi: University of Helsinki, Finland

Supplementary Table 2: The number of total reads, the percentage of mapped reads and the number of transcripts reconstructed by Cufflinks for each RNA-seq sample. *corresponds to punch biopsies from the nasal planum of Labrador Retrievers (Jagannathan, V. *et al*., 2013).

Species	Program	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC
Nematode	FEELnc shuffle	0.961	0.768	0.864	0.885	0.91	0.76
	CNCI	0.761	0.887	0.912	0.811	0.83	0.635
Chicken	CNCI	0.908	0.983	0.981	0.946	0.943	0.893
	FEELnc shuffle	0.945	0.836	0.852	0.891	0.896	0.786
Chimpanzee	CNCI	0.876	0.981	0.979	0.929	0.925	0.862
	FEELnc shuffle	0.946	0.897	0.902	0.922	0.923	0.844
Cow	CNCI	0.93	0.993	0.993	0.962	0.96	0.925
	FEELnc shuffle	0.96	0.942	0.943	0.951	0.952	0.903
Fly	FEELnc shuffle	0.974	0.763	0.804	0.868	0.881	0.754
	CNCI	0.94	0.728	0.776	0.834	0.85	0.684
Gorilla	FEELnc shuffle	0.947	0.953	0.953	0.95	0.95	0.9
	CNCI	0.866	0.994	0.994	0.925	0.925	0.859
Opossum	CNCI	0.845	0.977	0.973	0.913	0.905	0.832
	FEELnc shuffle	0.947	0.88	0.888	0.914	0.916	0.829
Orangutan	FEELnc shuffle	0.944	0.934	0.935	0.939	0.939	0.878
	CNCI	0.833	0.985	0.981	0.912	0.901	0.832
Platypus	CNCI	0.799	0.974	0.965	0.891	0.874	0.791
	FEELnc shuffle	0.884	0.861	0.864	0.872	0.874	0.745
Rat	CNCI	0.91	0.855	0.861	0.882	0.885	0.766
	FEELnc shuffle	0.96	0.751	0.794	0.856	0.869	0.727
Rhesus	FEELnc shuffle	0.944	0.937	0.937	0.94	0.941	0.881
	CNCI	0.853	0.988	0.986	0.92	0.914	0.849
Arabidopsis	CNCI	0.845	0.982	0.984	0.905	0.909	0.821
	FEELnc shuffle	0.988	0.787	0.865	0.904	0.922	0.809
Zebrafish	CNCI	0.911	0.924	0.922	0.917	0.917	0.835
	FEELnc shuffle	0.956	0.848	0.863	0.902	0.907	0.809

Supplementary Table 3: Performance metrics of FEELnc (*shuffle* mode) and CNCI tested on 5k NONCODE lncRNAs and 5k EnsEMBL mRNA annotations. Programs are ranked by MCC values per species tests. Species rows in bold indicate FEELnc MCC values higher than CNCI. FEELnc shuffle corresponds to the training of FEELnc with species-specific mRNAs (positive class) and species-specific shuffled mRNAs with preserved 7-mer frequencies (negative class).

Species	Abrevation	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC	Time of speciation (Myr)
Nematode	Cele	0.973	0.306	0.584	0.639	0.73	0.374	709
Arabidopsis	Atha	0.969	0.355	0.6	0.662	0.741	0.410	1434

Fly	Dmel	0.967	0.437	0.632	0.702	0.764	0.476	709
Cow	Btau	0.97	0.581	0.698	0.775	0.812	0.597	95
Zebrafish	Drer	0.942	0.711	0.765	0.826	0.844	0.671	340
Platypus	Oana	0.942	0.725	0.774	0.834	0.85	0.683	171
Mouse	Mmus	0.946	0.799	0.825	0.873	0.881	0.753	91.9
Gorilla	Ggor	0.945	0.81	0.832	0.878	0.885	0.762	8.9
Chicken	Ggal	0.93	0.845	0.857	0.887	0.892	0.778	292
Opossum	Mdom	0.931	0.849	0.86	0.89	0.894	0.782	158
Rat	Rnor	0.874	0.911	0.908	0.893	0.891	0.786	91.9
Orangutan	Pabe	0.932	0.874	0.881	0.903	0.906	0.807	15.8
Rhesus	Mmul	0.93	0.884	0.889	0.907	0.909	0.815	26
Chimpanzee	Pabe	0.918	0.905	0.906	0.911	0.912	0.823	6.7
Human	-	0.916	0.913	0.914	0.915	0.915	0.830	0

Supplementary Table 4: FEELnc performance with NONCODE lncRNAs as training set on the human HT set. Time of speciation data was extracted from (Hedges et al, MBE, 2015).

Features/biotypes		CanFam3.1	CanFam3.1-plus	New
Genes	mRNAs*	21,474	21,810	**336**
	lncRNAs**	8,008	10,444	**2,436**
Transcripts	mRNAs	100,110	158,750	**58,640**
	lncRNAs	12,506	22,880	**10,374**

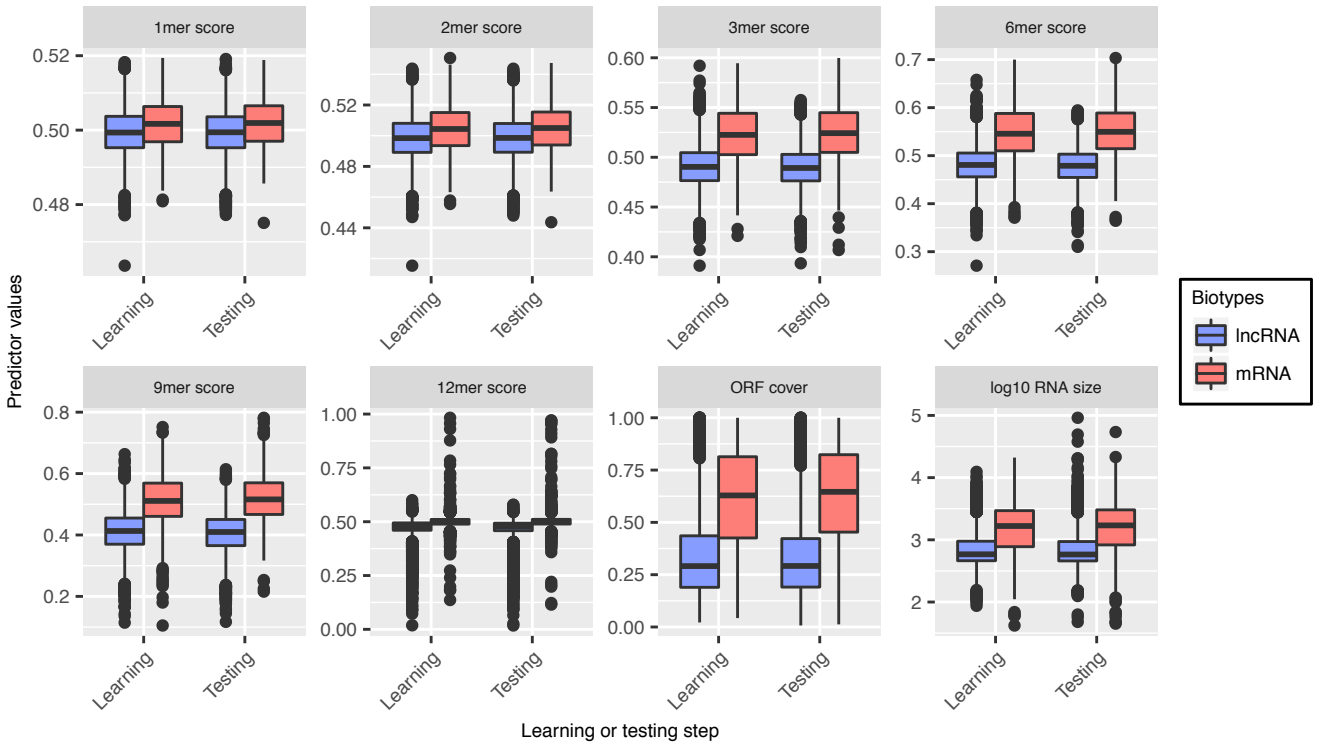
Supplementary Table 5: Comparison of CanFam3.1 versus CanFam3.1-plus annotations. Bold values correspond to the number of new genes/transcripts in CanFam3.1-plus. * Genes with at least one mRNA transcript; ** Genes with at least one lncRNA transcript (without any mRNA).

UTR and CDS comparison	CanFam3.1	CanFam3.1-plus	CanFam3.1	CanFam3.1-plus
	5'UTR + 3'UTR length		CDS length	
Transcripts with UTR or CDS	88,446	132,682	100,110	158,750
Median	1,194	3,033	1,188	1,215
Mean	2,301	3,783	1,652	1,680

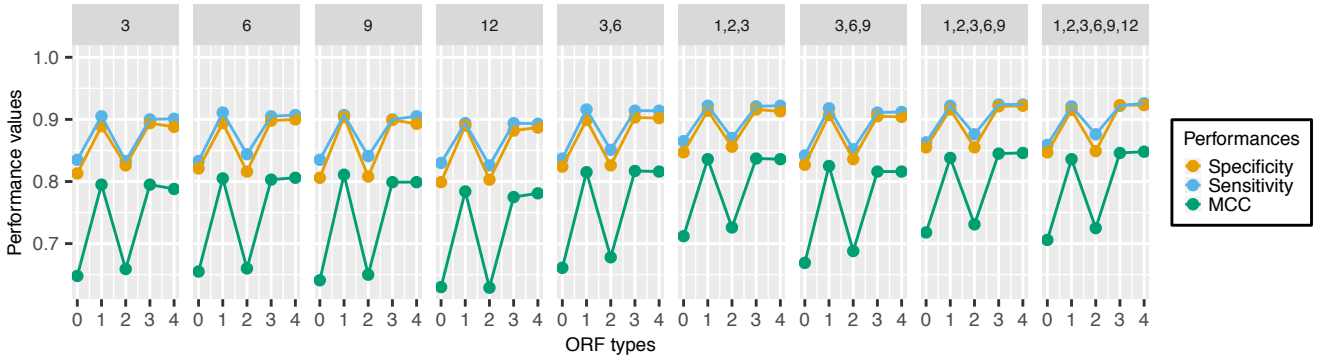
Supplementary Table 6: Comparison between the number of transcripts with an UTR and/or a CDS annotated and the length of these features between CanFam3.1 and the novel CanFam3.1-plus annotation.

Error types	FEELnc option	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC
Mutation	default	0.786	0.921	0.909	0.853	0.843	0.714
	wholeSeq	**0.869**	**0.749**	**0.776**	**0.809**	**0.82**	**0.622**
Deletion	default	0.234	0.943	0.803	0.589	0.363	0.251
	wholeSeq	**0.861**	**0.732**	**0.762**	**0.796**	**0.808**	**0.597**
Insertion	default	0.226	0.949	0.815	0.588	0.354	0.253
	wholeSeq	**0.818**	**0.807**	**0.809**	**0.813**	**0.814**	**0.625**

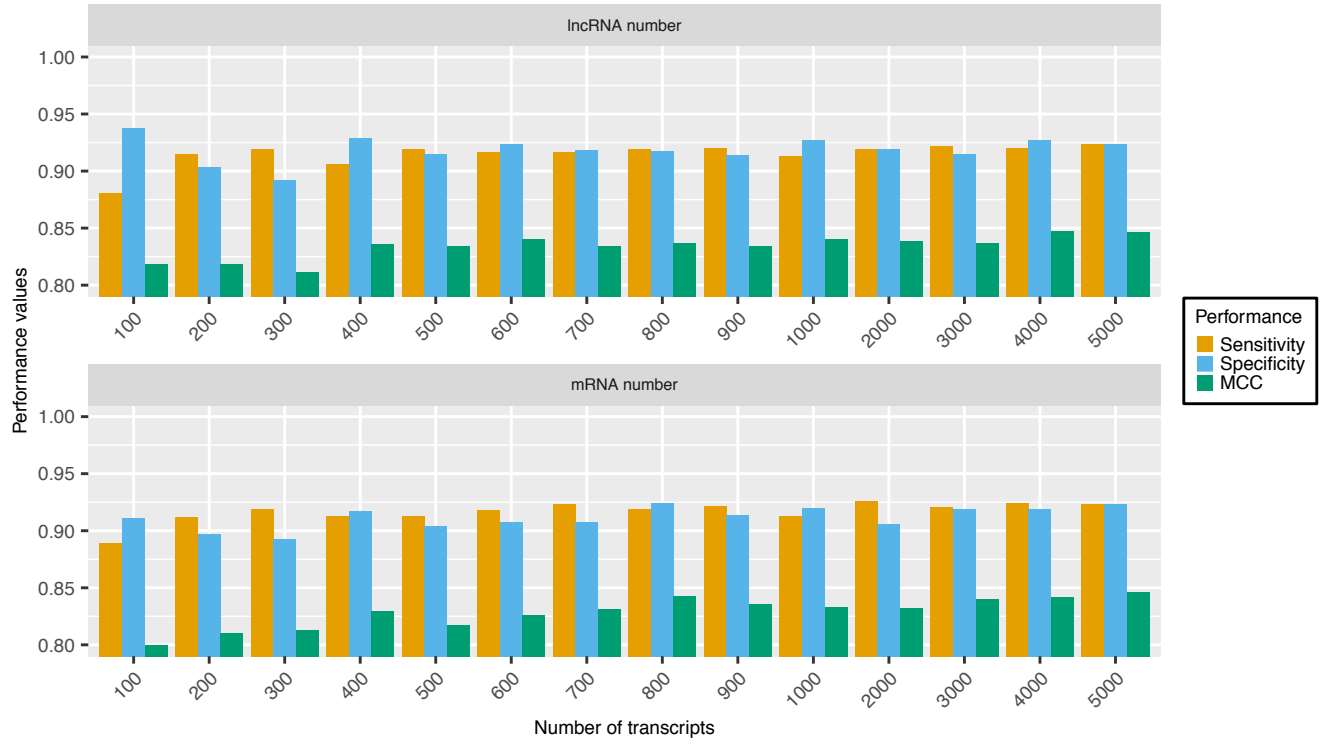
Supplementary Table 7: Results of FEELnc$_{codpot}$ (`--wholeSeq` option) on modified mRNA and lncRNA testing sequences. The errors in the human testing set (HT) were simulated as follows: for each sequence, a number of nucleotides between 5 and 15% have been mutated (Mutation), deleted (Deletion) or inserted (Insertion). FEELnc$_{codpot}$ has been run as follow: default: default options of FEELnc; wholeSeq: *k*-mer frequencies made on the whole transcript sequence and the 'ORF coverage' predictor has been removed from the random forest predictors list.



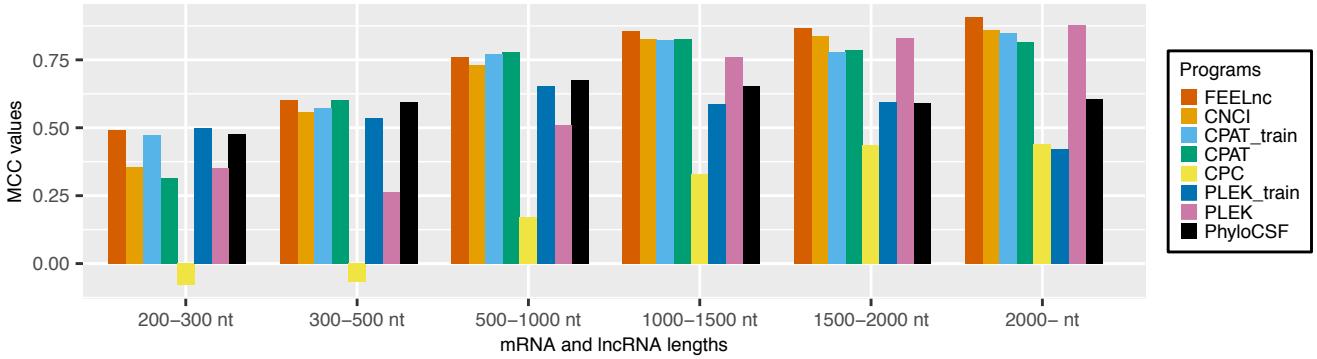
Supplementary Figure 1: Distribution of FEELnc predictor values on the human learning dataset (HT) composed of 5,000 lncRNAs and 5,000 mRNAs annotated in GENCODE v24.



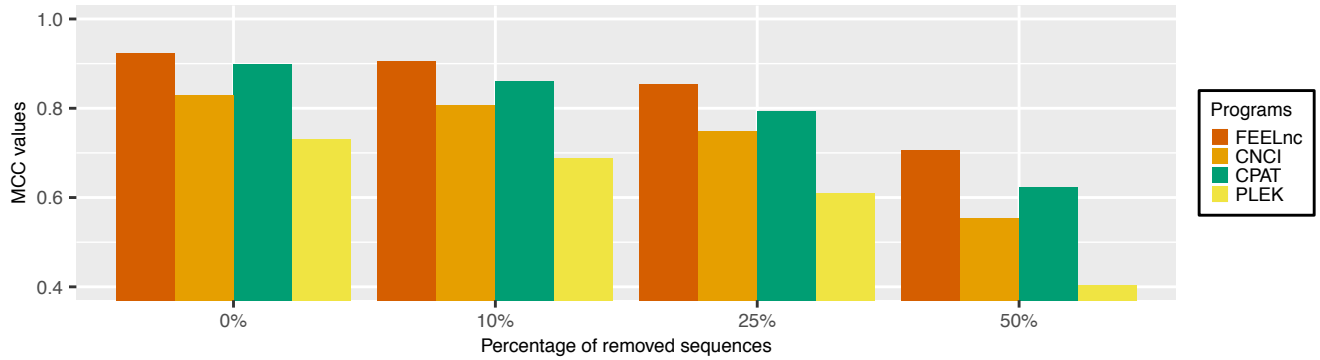
Supplementary Figure 2: FEELnc performance (sensitivity, specificity and MCC in orange, blue and green, respectively) with respect to different combination of ORF types (x-axis) and multi *k*-mer lists (grey panels) on the human dataset.



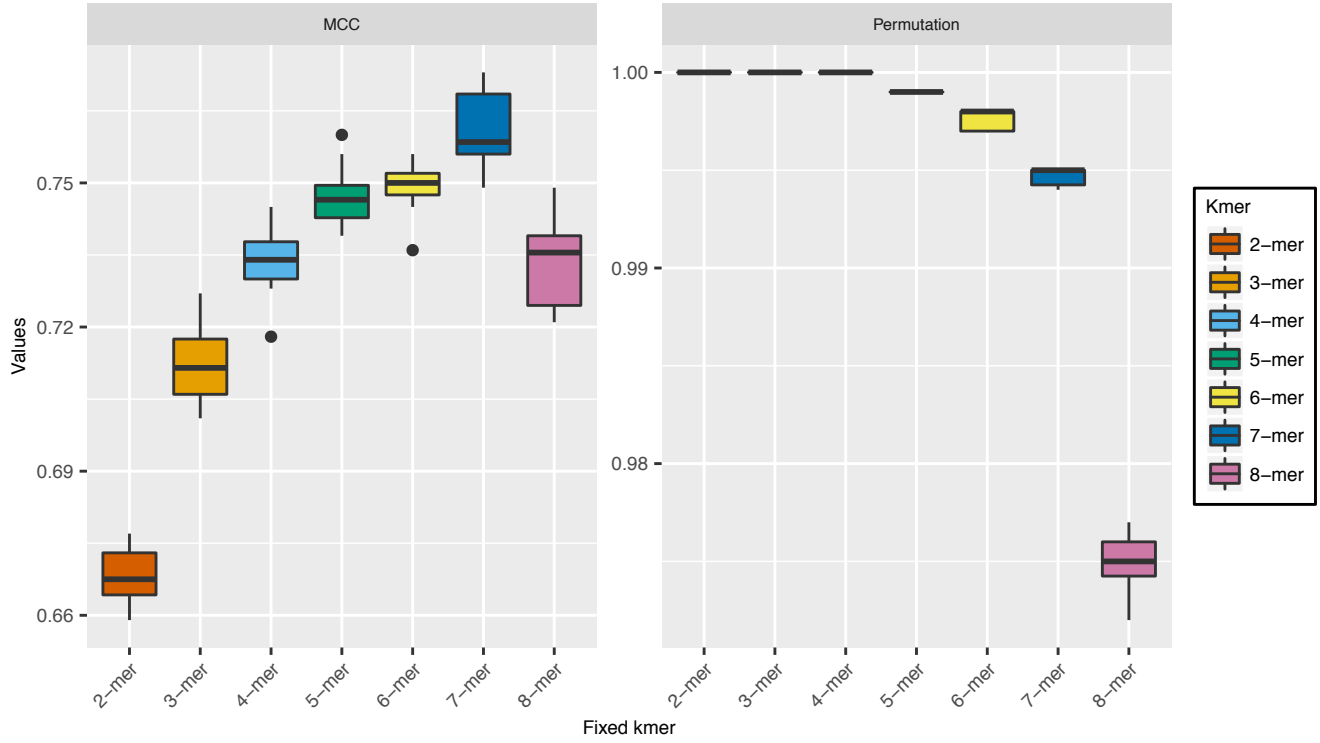
Supplementary Figure 3: FEELnc performance (sensitivity, specificity and MCC in orange, blue and green respectively) with respect to different numbers of learning transcripts in x-axis with lncRNAs (top panel) and mRNAs (bottom panel).



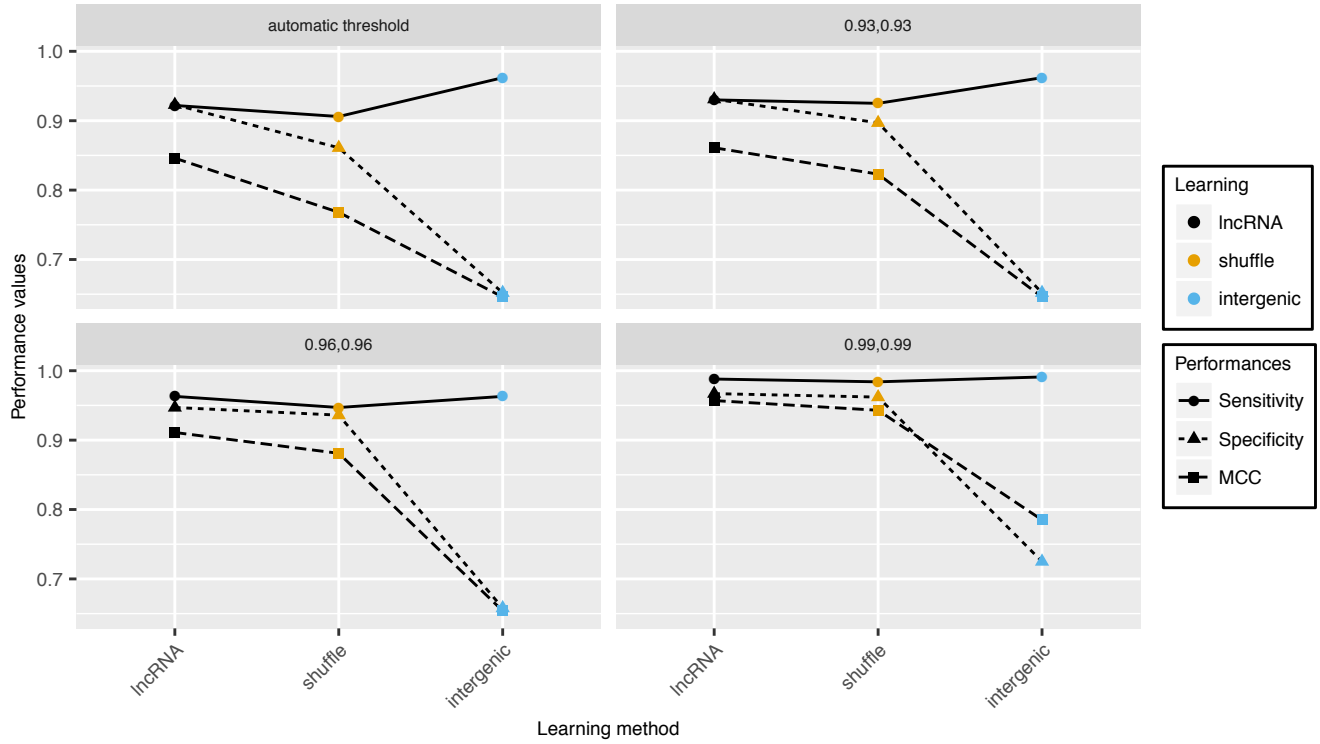
Supplementary Figure 4: Program benchmarking for different sizes of tested lncRNA and mRNA transcripts from the HT dataset.



Supplementary Figure 5: Program benchmarking for different percentages of shortened sequences in 5'-end or 3'-end for lncRNA and mRNA tested sequences from the HT dataset.



Supplementary Figure 6: FEELnc MCC value using the *shuffle* strategy on the HT dataset with different size of fixed *k*-mer frequencies by the UShuffle program (left panel). The corresponding ratio value of permutated/non-permutated sequences by Ushuffle (right panel).



Supplementary Figure 7: Benchmarking FEELnc performance (sensitivity, specificity, MCC in dot, triangle and square respectively) on the HT dataset for the 3 learning strategies: using human lncRNAs from the GENCODE HT set, shuffled mRNAs and intergenic human sequences, denoted *lncRNA*, *shuffle* and *intergenic* respectively (in black, orange and blue). Panels correspond to FEELnc with: automatic CPS threshold (top-left), 0.93 specificity thresholds for both lncRNAs/mRNAs (top-right), 0.96 specificity thresholds for both lncRNAs/mRNAs (bottom left) and 0.99 specificity thresholds for both lncRNAs/mRNAs (bottom right).